INFORMATION CLASSIFYING DEVICE

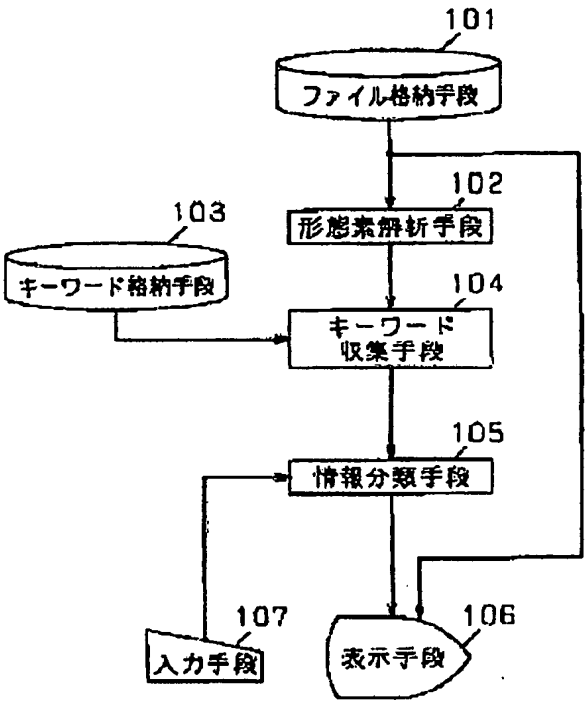
Patent number: JP10283366
Publication date: 1998-10-23

Inventor: MIZUTANI KENJI; OZAWA JUN; IMANAKA TAKESHI
Applicant: MATSUSHITA ELECTRIC IND CO LTD

Classification:
- international: G06F17/30
- european:
Application number: JP19970090656 19970409
Priority number(s): JP19970090656 19970409

Abstract of **JP10283366**

PROBLEM TO BE SOLVED: To automatically classify files containing texts by inputting the output of an information classifying means and the output of a file storage means and providing files for a user through a display means according to the classification result of the information classifying means. SOLUTION: A file containing a text consisting of character codes is stored in a file storage means 101 and a morpheme analyzing means 102 takes a morpheme analysis of a text part of the file in the file storage means 101 and outputs the result to a key word gathering means 104 together with the identifier of the file. The key word gathering means 104 gathers only key words stored in a key word storage means 103 from the result of the morpheme analysis and outputs them to the information classifying means 105 together with the identifier of the file. Then the information classifying means 105 classifies the identifier of the file with the key words attached thereto and displays and provides the file for the user through a display means 106 according to the classification result.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－283366

(43)公開日　平成10年(1998)10月23日

(51)Int.Cl.*	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３１０Ｄ
			15/40	３７０Ａ
			15/401	３１０Ａ

審査請求　未請求　請求項の数７　ＯＬ　（全　７　頁）

(21)出願番号　特願平9－90656

(22)出願日　平成９年(1997)４月９日

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　水谷　研治
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　小澤　順
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　今中　武
大阪府門真市大字門真1006番地　松下電器産業株式会社内

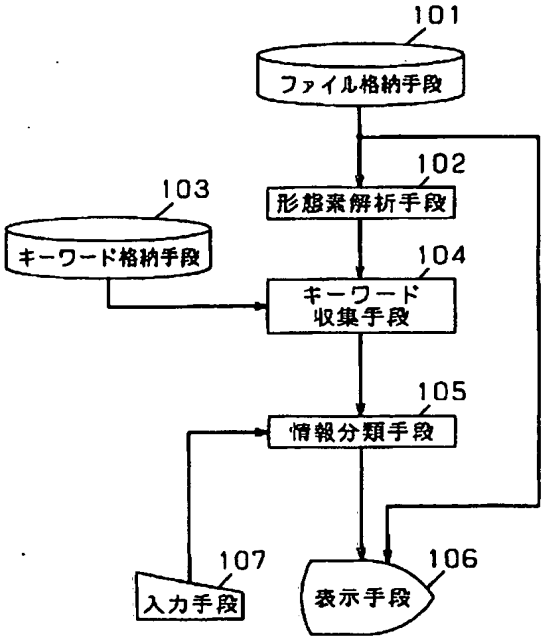
(74)代理人　弁理士　滝本　智之　（外１名）

(54)【発明の名称】　情報分類装置

(57)【要約】

【課題】　テキストを含むファイルを自動分類する装置の提供。

【解決手段】　ファイル格納手段１０１が出力するファイルのテキスト部分を形態素解析してファイルの識別子と共に出力する形態素解析手段１０２と、ファイルを分類するためのキーワードを格納するキーワード格納手段１０３と、キーワード格納手段１０３の出力と形態素解析手段１０２の出力とを入力として、形態素解析の結果の中からキーワードを収集してファイルの識別子と共に出力するキーワード収集手段１０４と、利用者が最近の分類結果と異なる分類結果を要求するための入力手段１０７と、入力手段１０７の出力とキーワード収集手段１０４の出力とを入力としてファイルの識別子をキーワードで分類して出力する情報分類手段１０５と、情報分類手段１０５の出力とファイル格納手段１０１の出力とを入力として、情報分類手段１０５が分類した結果を利用者に提供する。



【特許請求の範囲】
【請求項1】文字コードによって構成されるテキストを含むファイルを格納するファイル格納手段と、前記ファイル格納手段が出力するファイルのテキスト部分に対して形態素解析を行って前記ファイルの識別子と共に出力する形態素解析手段と、前記ファイルを分類するためのキーワードを格納するキーワード格納手段と、前記キーワード格納手段の出力と前記形態素解析手段の出力とを入力として、前記形態素解析の結果の中から前記キーワードだけを収集して前記ファイルの識別子と共に出力するキーワード収集手段と、利用者が最近の分類結果と異なる分類結果を要求するための入力手段と、前記入力手段の出力と前記キーワード収集手段の出力とを入力として前記ファイルの識別子を前記キーワードで分類して出力する情報分類手段と、前記情報分類手段の出力と前記ファイル格納手段の出力とを入力として、前記情報分類手段が分類した結果に従って前記ファイルを利用者に提供する表示手段によって構成される情報分類装置。
【請求項2】情報分類手段は、キーワードを持つファイルの識別子の集合と利用者からの指示とを入力として、ファイルを分類する前記キーワードを選択し、分類キーワード集合として出力する初期キーワード選択手段と、前記初期キーワード選択手段の出力を入力として、前記分類キーワード集合を洗練して出力する分類キーワード洗練手段と、前記分類キーワード洗練手段の出力を入力として、前記ファイルの識別子を前記分類キーワード集合に含まれる各キーワードに割り当てるファイル集合生成手段と、前記分類キーワード集合によって分類された前記ファイルの識別子の各集合を前記初期キーワード選択手段に出力して再帰的な分類を行わせる再帰的分類制御手段によって構成されることを特徴とする請求項1記載の情報分類装置。
【請求項3】初期キーワード選択手段は、利用者からの指示がなければ、キーワード収集手段が収集したキーワードを、前記キーワードが出現するファイルの数が多い順に一列に並べて、最上位から一定数の前記キーワードを分類キーワード集合として選択して出力することを特徴とする請求項1記載の情報分類装置。
【請求項4】初期キーワード選択手段は、利用者から最近の分類結果と異なる分類結果を要求する指示があれば、最近並べたキーワードの列について、最近選択した分類キーワード集合に含まれる前記キーワードを最下位に順序を保存して移動した後、最上位から一定数の前記キーワードを分類キーワード集合として選択して出力することを特徴とする請求項1記載の情報分類装置。
【請求項5】分類キーワード洗練手段は、分類キーワード集合に含まれるキーワードが出現するファイルの数を前記分類キーワード集合の評価関数として、前記分類キーワード集合に含まれる1つのキーワードを、まだ前記分類キーワード集合に含まれたことがない1つのキーワードと置換し、前記評価関数の値が前記置換の直前の値より増加する限り前記置換を行って、前記分類キーワード集合を更新することを特徴とする請求項1記載の情報分類装置。
【請求項6】ファイル集合生成手段は、分類キーワード集合に含まれるキーワードを、前記キーワードが出現するファイルの数が多い順に一列に並べて、前記キーワードに割り当てるファイルの識別子の集合を、前記キーワードよりも下位のキーワードが1つも出現しない前記ファイルの識別子に限定し、かつ前記分類キーワード集合に含まれるキーワードが1つも出現しないファイルについては、その他を意味する特殊キーワードを前記分類キーワード集合に追加して、前記特殊キーワードに前記ファイルの識別子を割り当てることを特徴とする請求項1記載の情報分類装置。
【請求項7】文字コードによって構成されるテキストを含むファイルをキーワードを付けて分類するプログラム製品であり、以下のステップを実現するプログラム記録媒体を含む：ファイルに含まれる文字コードによって構成されるテキスト部分を形態素解析するステップ、
前記形態素解析の結果から前記ファイルを分類するために使用するキーワードを収集して、ファイルの識別子と共に出力するステップ、
キーワードを持つファイルの識別子の集合を、前記キーワードで分類するステップ。
【発明の詳細な説明】
【0001】
【発明の属する技術分野】本発明は、文字コードによって構成されるテキストを含むファイルをキーワードを付けて分類する装置に関するものである。
【0002】
【従来の技術】データベース・システムは一般に、検索目的を持った利用者が目的のファイルに容易に到達できるように、キーワード論理式などを入力するインタフェースを用意している。しかしながら、特に検索目的を持たず、データベースの中にどのようなファイルが収納されているかということに興味を持つ利用者にとっては、このようなインタフェースはあまり役に立たない。データベースの内容の一覧を提供するために、従来は、データベースの管理者があらかじめ固定的な概念体系を用意して、新しく追加するファイルの内容を理解してその体系における位置を決定したり、ファイルの提供者が位置を指定したり、あるいは、すでに手作業で分類したファイルとキーワードを比較して最も近い位置に自動分類して、利用者に分類結果を提供していた。
【0003】
【発明が解決しようとする課題】前述のあらかじめ固定的な概念体系を用意する方法では、新しい概念を持ったファイルが出現したときに利用者にその存在が伝わらないという問題が生じる。自動分類では、すでに分類され

ているファイルとキーワードが1つでも一致すれば概念的に近いと判断されて既存の概念に分類されるだけである。したがって、適当な時期に概念体系を修正して分類をやり直す必要があるが、その作業はデータベースの規模に比例して膨大な量になる。

【0004】本発明は、固定的な概念体系を利用するのではなく、ファイルに含まれるキーワードを概念として利用し、ファイルをそれに属する集合として自動分類して、ユーザにデータベースの内容の一覧を提供することを目的とする。

【0005】

【課題を解決するための手段】請求項1記載の本発明は、文字コードによって構成されるテキストを含むファイルを格納するファイル格納手段と、前記ファイル格納手段が出力するファイルのテキスト部分に対して形態素解析を行って前記ファイルの識別子と共に出力する形態素解析手段と、前記ファイルを分類するためのキーワードを格納するキーワード格納手段と、前記キーワード格納手段の出力と前記形態素解析手段の出力とを入力として、前記形態素解析の結果の中から前記キーワードだけを収集して前記ファイルの識別子と共に出力するキーワード収集手段と、利用者が最近の分類結果と異なる分類結果を要求するための入力手段と、前記入力手段の出力と前記キーワード収集手段の出力とを入力として前記ファイルの識別子を前記キーワードで分類して出力する情報分類手段と、前記情報分類手段の出力と前記ファイル格納手段の出力とを入力として、前記情報分類手段が分類した結果に従って前記ファイルを利用者に提供する表示手段によって構成される情報分類装置である。

【0006】請求項2記載の本発明は、情報分類手段が、キーワードを持つファイルの識別子の集合と利用者からの指示とを入力として、ファイルを分類する前記キーワードを選択し、分類キーワード集合として出力する初期キーワード選択手段と、前記初期キーワード選択手段の出力を入力として、前記分類キーワード集合を洗練して出力する分類キーワード洗練手段と、前記分類キーワード洗練手段の出力を入力として、前記ファイルの識別子を前記分類キーワード集合に含まれる各キーワードに割り当てるファイル集合生成手段と、前記分類キーワード集合によって分類された前記ファイルの識別子の各集合を前記初期キーワード選択手段に出力して再帰的な分類を行わせる再帰的分類制御手段によって構成される情報分類装置である。

【0007】請求項3記載の本発明は、初期キーワード選択手段が、利用者からの指示がなければ、キーワード収集手段が収集したキーワードを、前記キーワードが出現するファイルの数が多い順に一列に並べて、最上位から一定数の前記キーワードを分類キーワード集合として選択して出力する情報分類装置である。

【0008】請求項4記載の本発明は、初期キーワード選択手段が、利用者から最近の分類結果と異なる分類結果を要求する指示があれば、最近並べたキーワードの列について、最近選択した分類キーワード集合に含まれる前記キーワードを最下位に順序を保存して移動した後、最上位から一定数の前記キーワードを分類キーワード集合として選択して出力する情報分類装置である。

【0009】請求項5記載の本発明は、分類キーワード洗練手段が、分類キーワード集合に含まれるキーワードが出現するファイルの数を前記分類キーワード集合の評価関数として、前記分類キーワード集合に含まれる1つのキーワードを、まだ前記分類キーワード集合に含まれたことがない1つのキーワードと置換し、前記評価関数の値が前記置換の直前の値より増加する限り前記置換を行って、前記分類キーワード集合を更新する情報分類装置である。

【0010】請求項6記載の本発明は、ファイル集合生成手段が、分類キーワード集合に含まれるキーワードを、前記キーワードが出現するファイルの数が多い順に一列に並べて、前記キーワードに割り当てるファイルの識別子の集合を、前記キーワードよりも下位のキーワードが1つも出現しない前記ファイルの識別子に限定し、かつ前記分類キーワード集合に含まれるキーワードが1つも出現しないファイルについては、その他を意味する特殊キーワードを前記分類キーワード集合に追加して、前記特殊キーワードに前記ファイルの識別子を割り当てる情報分類装置である。

【0011】

【発明の実施の形態】本発明の一実施の形態の情報分類装置全体の構成を表すブロック図を図1に示す。ファイル格納手段101は、文字コードによって構成されるテキストを含むファイルを格納する。形態素解析手段102は、ファイル格納手段101のファイルのテキスト部分に対して形態素解析を行ってファイルの識別子と共に出力する。キーワード格納手段103は、ファイルを分類するためのキーワードを格納する。キーワード収集手段104は、形態素解析の結果の中からキーワード格納手段103に格納されているキーワードだけを収集して、ファイルの識別子と共に出力する。情報分類手段105は、ファイルの識別子をそれに付随するキーワードによって分類する。表示手段106は、ファイル格納手段101のファイルの内容を分類結果に従って利用者に提供する。入力手段107は、利用者が提供された分類結果と異なる分類を希望するときに、情報分類手段105にその要求を伝える。

【0012】次に本実施の形態の動作を説明する。例として、図2に示すラーメンの飲食店について記述した5つのファイルがファイル格納装置101に格納されているとする。それぞれのファイルの識別子は、file1, file2, file3, file4, file5である。

【0013】形態素解析手段102は、ファイル格納手

段101に格納されているファイルのテキスト部分について形態素解析を行い、ファイルの識別子と共に出力する。図2に示すファイルについて、形態素解析手段102が処理した結果の、名詞のみを取り出した結果を図3に示す。

【0014】キーワード格納手段103には、分類に使用するキーワードを列挙する。例を図4に示す。

【0015】キーワード収集手段104は、形態素解析手段102の出力の中から、キーワード格納手段103に格納されている単語だけを取り出して、ファイルの識別子と共に出力する。図3の形態素解析の結果を、キーワード収集手段104が処理した結果を図5に示す。

【0016】情報分類手段105は、キーワード収集手段が104が出力するキーワードを持つファイルの識別子の集合を、キーワードで分類して出力する。情報分類手段105の詳細な構成を示すブロック図を図6に示す。

【0017】初期キーワード選択手段601は、キーワードをそれが出現するファイル数が多い順に並べ、最上位から一定数のキーワードを分類キーワード集合として選択する。図5のキーワード収集手段104の出力を、キーワードを横軸として出現ファイル数が多い順に左から並べた結果を図7に示す。分類キーワード集合として選択するキーワードの数を2とすると、分類キーワードの集合は、｛ラーメン、しょうゆ味｝となる。

【0018】分類キーワード洗練手段602は、初期キーワード選択手段601が出力する分類キーワード集合に含まれるキーワードが、より多くのファイルに出現するように他のキーワードと置換する。まず、分類キーワード集合に含まれるキーワードが出現するファイルの数を評価関数とする。そして、分類キーワード集合に含まれる1つのキーワードを、まだ分類キーワード集合に含まれたことがないキーワードと置換する操作を、評価関数の値が増加する限り繰り返す。図7の例で、分類キーワード集合が、

｛ラーメン、しょうゆ味｝

に設定されているとき、評価関数の値は4である。分類集合に含まれるキーワードの「ラーメン」と「しょうゆ味」を、まだ分類集合に含まれたことがないキーワードの「焼き豚」と置換し、評価関数の値を計算するといずれの場合も4である。したがって、評価関数の値が増加しないので、分類キーワード洗練手段602は分類キーワード集合を、

｛ラーメン、しょうゆ味｝

として出力する。

【0019】ファイル集合生成手段603は、分類キーワード洗練手段602が出力する分類キーワード集合に従って、ファイルの識別子を分類する。まず、分類キーワード集合に含まれるキーワードを、それが出現するファイル数が多い順に並べて、キーワードに割り当てるファイルの識別子の集合を、そのキーワードよりも下位のキーワードが1つも出現しないファイルの識別子に限定する。図7の例で分類キーワード集合が、

｛ラーメン、しょうゆ味｝

であれば、キーワード「ラーメン」が出現するファイルは、

{file1, file3, file5}

であるが、file1にはそれよりも下位のキーワード「しょうゆ味」が出現するので、各キーワードに割り当てるファイルの識別子の集合は、

ラーメン：{file3, file5}

しょうゆ味：{file1, file4}

となる。また、分類キーワード集合に含まれるキーワードが1つも出現しないファイルについては、特殊キーワード「その他」を分類キーワード集合に追加し、それにファイルの識別子を割り当てる。図7の例では、

その他：{file2}

となり、ファイル集合生成手段603から情報分類手段105の結果として、

｛ラーメン：{file3, file5}、しょうゆ味：{file1, file4}、その他：{file2}｝

が出力される。

【0020】再帰的分類制御手段604は、ファイル集合生成手段603が分類した結果をさらに細分類するときに使用する。すなわち、すでに分類されたファイルの識別子とそのファイルに出現するキーワードの集合を初期キーワード選択手段601に与えることで、分類されたファイルの識別子をさらにキーワードで分類する。

【0021】表示手段106は、情報分類手段105の結果を木構造に変換して、利用者にデータベースの内容の一覧を提供する。情報分類手段105の出力が、

｛ラーメン：{file3, file5}、しょうゆ味：{file1, file4}、その他：{file2}｝

のときは、図8に示すような出力結果が得られる。利用者は、この出力結果を見て、他の分類結果を要求したいときに入力手段107を用いる。入力手段107は情報分類手段105に接続され、初期キーワード選択手段201にその要求が伝えられる。

【0022】初期キーワード選択手段601は、キーワード収集手段104が出力した結果から最近選択した分類キーワード集合を記憶している。入力手段107から利用者の要求が伝えられると、最近並べたキーワードの列について、最近選択した分類キーワードに含まれるキーワードの列を、最下位に順序を保存して移動した後、最上位から一定数のキーワードを分類キーワード集合として選択して出力する。図7の例では、分類キーワードとして

｛ラーメン、しょうゆ味｝

を最近選択したので、それを順序を保存して最下位のキーワード「焼き豚」の次に移動し、図9のようなキーワ

ードの列を作る。そして最上位から2つのキーワードを選択して、分類キーワード集合、
{焼き豚、ラーメン}
を選択して出力する。分類キーワード洗練装置602以降の処理は同様であり、情報分類装置の出力として、
{ラーメン：{file1, file5}、焼き豚：{file2, file3}、その他：{file4}}
が出力される。表示手段106には、前回の図8の分類結果とは異なる、図10に示すようなデータベースの内容の一覧が利用者に提供される。

【0023】なお、本発明は文字コードによって構成されるテキストを含むファイルであればどのような種類のファイルでも分類することができる。ファイルをインターネット上のホームページを構成するHTMLファイル、ファイルの識別子をそのURLアドレスとすれば、本発明の情報分類装置をホームページの分類システムとして利用することができる。

【0024】

【発明の効果】以上述べたところから明らかなように、本発明は、キーワードを概念として利用し、文字コードによって構成されるテキストを含むファイルを自動分類するので、新しい概念を持ったファイルが出現しても、キーワードを保守するだけで容易に概念体系の更新が可能であり、利用者にデータベースの内容の一覧を迅速に提供できるという長所を有する。

【図面の簡単な説明】

【図1】本発明の一実施の形態の情報分類装置の全体の構成を表すブロック図

【図2】同実施の形態の動作を説明するための図1のファイル格納手段101の一例を示す図

【図3】同実施の形態の動作を説明するための図1の形態素解析手段102の出力の一例を示す図

【図4】同実施の形態の動作を説明するための図1のキーワード格納手段103の一例を示す図

【図5】同実施の形態の動作を説明するための図1のキーワード収集手段104の出力の一例を示す図

【図6】同実施の形態の動作を説明するための図1の情報分類手段105の詳細なブロック図

【図7】同実施の形態の動作を説明するための図6の初期キーワード選択手段601の内部状態の一例を示す図

【図8】同実施の形態の動作を説明するための図1の表示手段106の一例を示す図

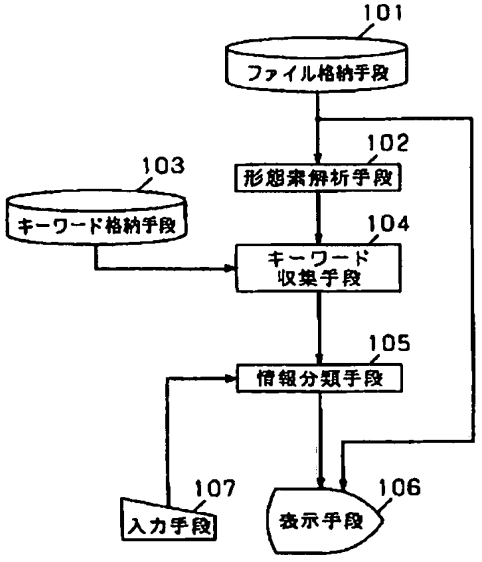
【図9】同実施の形態の動作を説明するための図6の初期キーワード選択手段601の内部状態の一例を示す図

【図10】同実施の形態の動作を説明するための図1の表示手段106の一例を示す図

【符号の説明】

101　ファイル格納手段
102　形態素解析手段
103　キーワード格納手段
104　キーワード収集手段
105　情報分類手段
106　表示手段
107　入力手段
601　初期キーワード選択手段
602　分類キーワード洗練手段
603　ファイル集合生成手段
604　再帰的分類制御手段

【図1】



【図3】

ファイルの識別子	形態素解析の結果　（名詞のみ）
file1	ラーメン、店、しょうゆ味、東京、一番
file2	焼き豚、九州、博多、味
file3	札幌、市内、ラーメン、焼き豚、駅
file4	長崎、しょうゆ味、伝統、味、秘伝、調合
file5	ラーメン、屋台、店主、夜間、営業

【図2】

file1
<ラーメンの店>
東京で一番のしょうゆ味ラーメンです。

file2
焼き豚にこだわりました。
九州博多の味をそのままあなたに
お届けします。

file3
札幌市内、駅からすぐです。
焼き豚がいっぱいのったラーメンをどうぞ。

file4
<長崎の伝統の味>
秘伝の調合でしょうゆ味を作りました。

file5
「なつかしい屋台のラーメン」
※店主は夜間しか営業いたしません。

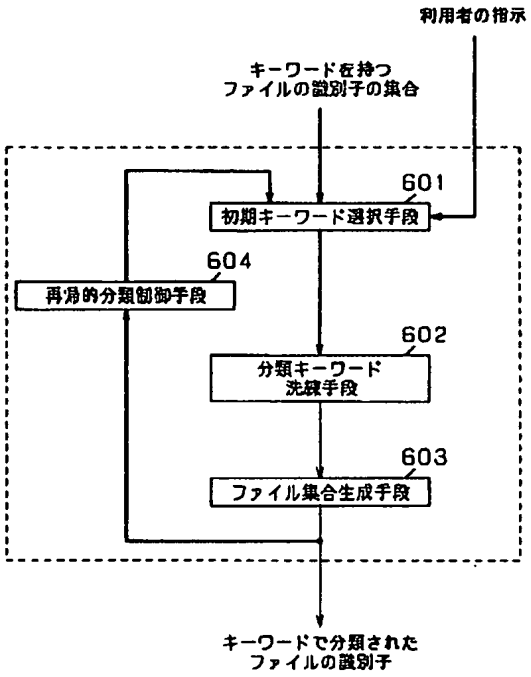
【図4】

ラーメン
メンマ
ネギ
焼き豚
もやし
しょうゆ味
みそ味
玉子
塩味
海苔
ほうれん草
スープ
にんにく
ラー油
とんこつ

【図5】

ファイルの識別子	分類キーワードの候補
file1	ラーメン、しょうゆ味
file2	焼き豚
file3	ラーメン、焼き豚
file4	しょうゆ味
file5	ラーメン

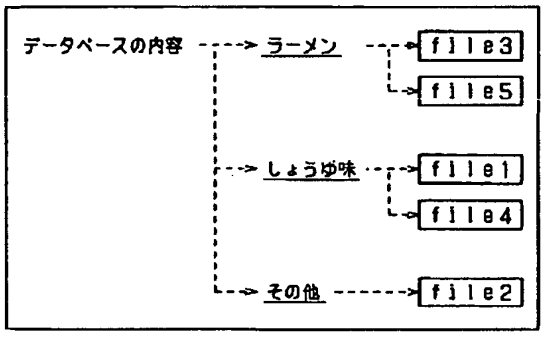
【図6】



【図7】

キーワード	ラーメン	しょうゆ味	焼き豚
出現ファイル数 / ファイルの識別子	3	2	2
file1	存在	存在	
file2			存在
file3	存在		存在
file4		存在	
file5	存在		

【図8】



【図9】

キーワード	焼き豚	ラーメン	しょうゆ味
出現ファイル数 / ファイルの識別子	2	3	2
file1		存在	存在
file2	存在		
file3	存在	存在	
file4			存在
file5		存在	

【図10】